Check for updates

Article

Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–30
DOI: 10.3102/1076998619874089
Article reuse guidelines: sagepub.com/journals-permissions
© 2019 AERA. http://jebs.aera.net

Validation Methods for Aggregate-Level Test Scale Linking: A Case Study Mapping School District Test Score Distributions to a Common Scale

Sean F. Reardon
Demetra Kalogrides
Stanford University

Andrew D. Ho
Harvard Graduate School of Education

Linking score scales across different tests is considered speculative and fraught, even at the aggregate level. We introduce and illustrate validation methods for aggregate linkages, using the challenge of linking U.S. school district average test scores across states as a motivating example. We show that aggregate linkages can be validated both directly and indirectly under certain conditions such as when the scores for at least some target units (districts) are available on a common test (e.g., the National Assessment of Educational Progress). We introduce precision-adjusted random effects models to estimate linking error, for populations and for subpopulations, for averages and for progress over time. These models allow us to distinguish linking error from sampling variability and illustrate how linking error plays a larger role in aggregates with smaller sample sizes. Assuming that target districts generalize to the full population of districts, we can show that standard errors for district means are generally less than .2 standard deviation units, leading to reliabilities above .7 for roughly 90% of districts. We also show how sources of imprecision and linking error contribute to both within- and between-state district comparisons within versus between states. This approach is applicable whenever the essential counterfactual question—"what would means/variance/progress for the aggregate units be, had students taken the other test?"—can be answered directly for at least some of the units.

Keywords: *linking; scaling; multilevel modeling; achievement testing; NAEP*

Introduction

As educational testing programs proliferate, nonoverlapping populations and incomparable scales can limit the scope of research about the correlates and causes of educational achievement. Linking is the psychometric solution to this

problem. Common persons, common populations, or common items across tests form the basis for estimated linking functions (Kolen & Brennan, 2014). These functions can enable mappings of scores from various tests to a common scale, enabling large-scale research about educational achievement. However, the bases for these linkages—common persons, populations, or items—are not always available at a large scale. When they are available, methods for evaluating the linkage for the purpose of large-scale research, rather than student-level uses like diagnosis and selection, are still in development (Thissen, 2007).

Dorans and Holland (2000) outline five requirements for equating (1) equal constructs, (2) equal reliability, (3) examinee indifference between tests, and (4) a symmetrical linking function that is (5) invariant across populations. These requirements are only realistically met within testing programs, not across them. For linkages that do not meet the stringent conditions of equating, the appropriateness of the linkage becomes dependent on the interpretations and uses of the linked scores.

We present a case of aggregate-level linking whose purpose is to support education research. First, we show how a common assessment at one level of aggregation (the state, in our example) can serve as the basis for a common-population linkage. Second, we demonstrate how the assessment can directly validate the linkage on which it is based, if the assessment also reports scores at a lower level of aggregation (the school district, here). Third, we show how to validate inferences about progress over time in addition to inferences about relative achievement. Fourth, we show how additional assessments that are common across a subset of the lower level units can provide indirect validation of the linking. Although none of the methods we present are new on its own, the logic and methods in this validation approach is likely to be useful in other aggregate linking scenarios.

A Case Comparing U.S. School District Achievement Across States

To understand how a "patchwork" administration of tests can support aggregate linking, we present the case of linking U.S. school district average scores to a common scale. U.S. school districts differ dramatically in their socioeconomic and demographic characteristics (Reardon, Yun, & Eitle, 1999; Stroub & Richards, 2013), and districts have considerable influence over instructional and organizational practices that may affect academic achievement (Whitehurst, Chingos, & Gallaher, 2013). Nonetheless, we have relatively little rigorous large-scale research describing national patterns of variation in achievement across districts, let alone an understanding of the factors that cause this variation. Such analyses generally require district-level test score distributions that are comparable across states. No such nationwide, district-level achievement data set currently exists because school districts do not administer a common set of assessments to all districts across states.

Existing assessments enable some comparisons of academic performance across states or school districts, but none provides comprehensive comparisons across grades, years, and all school districts. At the highest level, the National Assessment of Educational Progress (NAEP) provides comparable state-level scores in odd years, in reading and mathematics, in Grades 4 and 8. NAEP also provides district-level scores, but only for a small number of large urban districts under the Trial Urban District Assessment (TUDA) initiative: TUDA began with 6 districts in 2002 and slowly grew to 27 districts by 2017. Within individual states, we can compare district achievement within a given grade and year using state math and reading/English language arts (ELA) tests federally mandated by the No Child Left Behind act, administered annually in Grades 3 through 8. Comparing academic achievement across state lines requires either that districts administer a common test or that the scores on the state tests can be linked to a common scale. However, state accountability tests generally differ across states. Each state develops and administers its own tests; these tests may assess somewhat different content domains; scores are reported on different, state-determined scales; and proficiency thresholds are set at different levels of achievement. Moreover, the content, scoring, and definition of proficiency may vary within any given state over time and across grades.

As a result, direct comparisons of average scores or percentages of proficient students across states (or in many cases within states, across grades and years) are unwarranted and misleading. Average scores may differ because scales differ and because performance differs. Proficiency rates may differ because proficiency thresholds differ (Bandeira de Mello, Blankenship, & McLaughlin, 2009; Braun & Qian, 2007) and because performance differs. The ongoing rollout of common assessments developed by multistate assessment consortia (such as the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium) is certainly increasing comparability across states, but only to the extent that states use these assessments. Customization of content standards by states may also discourage the reporting of results on a common scale across states (Gewertz, 2015; U.S. Department of Education, 2009). Given the incomplete, divided, and declining state participation in these consortia, comprehensive, directly comparable district-level test score data in the United States remain unavailable.

In some cases, districts also administer voluntarily chosen assessments, often for lower stakes purposes. When two districts adopt the same such assessments, we can compare test scores on these assessments among districts. One of the most widely used assessments, the Measures of Academic Progress (MAP) test from Northwest Evaluation Association (NWEA), is voluntarily administered in several thousand school districts, over 20% of all districts in the country. However, the districts using MAP are neither a representative nor comprehensive sample of districts.

In this article, we present a validation strategy for comparisons of district-level test scores across states, years, and grades. We rely on a combination of (a) population-level state test score data from NAEP and state tests, (b) linear transformations that link state test scores to observed and interpolated NAEP scales, and (c) a set of validation checks to assess the accuracy of the resulting linked estimates. In addition, we provide formulas to quantify the uncertainty in both between- and within-state comparisons. Together, this represents a suite of approaches for constructing and evaluating linked estimates of test score distributions.

We use data from the ED*Facts* Initiative (U.S. Department of Education, 2015), NAEP, and NWEA. We obtain population-level state testing data from ED*Facts*; these data include counts of students in ordered proficiency categories for each district–grade–year–subject combination. We fit heteroskedastic ordered probit (HETOP) models to these district proficiency counts, resulting in estimated district means and variances on a state standardized (zero mean and unit variance) scale (Reardon, Shear, Castellano, & Ho, 2016). We then apply linear linking methods that adjust for test reliability (reviewed by Kolen & Brennan, 2014) to place each district's estimated score distribution parameters on a common national scale. Our linking methods are similar to those that Hanushek and Woessman (2012) used to compare country-level performance internationally. At the district level (Greene & McGee, 2011) and school level (Greene & Mills, 2014), the *Global Report Card* maps scores onto a national scale using proficiency rates, using a somewhat different approach than ours.[1] What we add to these standard linear linking methods are direct and indirect validation methods that take advantage of patchwork reporting of test scores at the target levels of aggregation. We also develop an approach to assessing the uncertainty in linked estimates resulting from both measurement error and linking error.

Although some have argued that using NAEP as a basis for linking state accountability tests is both infeasible and inappropriate for high-stakes student-level reporting (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999), our goal here is different. We do not attempt to estimate student-level scores, and we do not intend the results to be used for high-stakes accountability. Rather, our goal is to estimate transformations that render aggregate test score distributions roughly comparable across districts in different states so that the resulting district-level distributions can be used in aggregate-level research. We grant that NAEP and state tests may differ in many respects including content, testing dates, motivation, accommodations for language, accommodations for disabilities, and test-specific preparation. While accepting these sources of possible linking error, we focus on the counterfactual question that linking asks: How well do our linked district scores from state tests recover the average NAEP scores that these districts would have received had their students taken NAEP? In this way, we treat the issue of feasibility empirically, by using validation

checks to assess the extent to which our methods yield unbiased estimates of aggregate parameters of interest.

In any situation where states have some subset of districts with common scores on an alternative test, it is unlikely that those districts will be representative given their selection into the alternative test. When districts are representative, the methods we introduce will provide estimates of linking errors for the full population of districts. When districts are not representative, these methods provide estimates of linking errors for the particular subpopulations of districts they represent. In these latter cases, these methods provide evidence about the validity of the linkage for a subpopulation, a recommended step in the validation of any linking (Dorans & Holland, 2000).

Data

We use state accountability test score data and state NAEP data to link scores, and we use NAEP TUDA data and NWEA MAP data to evaluate the linkage. Under the ED*Facts* Initiative (U.S. Department of Education, 2015), states report frequencies of students scoring in each of several ordered proficiency categories for each tested school, grade, and subject (mathematics and reading/ELA). The numbers of ordered proficiency categories vary by state, from 2 to 5, most commonly 4. We use ED*Facts* data from 2009 to 2013, in Grades 3 through 8, provided to us by the National Center for Education Statistics under a restricted data use license. These data are not suppressed and have no minimum cell size. We also use reliability estimates collected from state technical manuals and reports for these same years and grades.[2]

Average NAEP scores and their standard deviations are reported for states and participating TUDA districts in odd years, in Grades 4 and 8, in reading and mathematics. In each state and TUDA district, these scores are based on an administration of the NAEP assessments to representative samples of students in the relevant grades and years. We use years 2009, 2011, and 2013 as a basis for linking, which include 17 TUDA districts in 2009 and 20 TUDA districts in 2011 and 2013.[3] The NAEP state and district means and standard deviations, as well as their standard errors, are available from the NAEP Data Explorer (U.S. Department of Education, n.d.). To account for NAEP initiatives to expand and standardize inclusion of English learners and students with disabilities over this time period, we rely on the Expanded Population Estimates (EPE) of means and standard deviations provided by the National Center of Education Statistics (see Braun, Zhang, & Vezzu, 2010; McLaughlin, 2005; National Institute of Statistical Sciences, 2009).[4]

Finally, we use data from the NWEA MAP test that overlaps with the years, grades, and subjects available in the ED*Facts* data: 2009 through 2013, Grades 3 through 8, and in reading/ELA and mathematics. Student-level MAP test score data (scale scores) were provided to us through a restricted-use data-sharing

agreement with NWEA. Several thousand school districts chose to administer the MAP assessment in some or all years and grades that overlap with our ED*Facts* data. Participation in the NWEA MAP is generally binary in districts administering the MAP; that is, in participating districts, either very few students or essentially all students are assessed. We exclude cases in any district's grade, subject, and year, where the ratio of assessed students to enrolled students is lower than 0.9 or greater than 1.1. This eliminates districts with scattered classroom-level implementation as well as very small districts with accounting anomalies. Excluded districts comprise roughly 10% of the districts using the NWEA MAP tests. After these exclusions, we estimate district–grade–subject–year means and standard deviations from student-level data reported on the continuous MAP scale.

Linking Methods

The first step in linking the state test scores to the NAEP scale is to estimate district-level score means and standard deviations and corresponding standard errors. If individual scale score data or district-level means and standard deviations were available, one could simply use these to obtain the necessary district-level parameter estimates. For this case study, such data were not readily available in most states. Instead, we estimate the district score means and standard deviations from the coarsened proficiency count data available in ED*Facts*, using the methods described in detail by Reardon, Shear, Castellano, and Ho (2016). These parameters are scaled relative to the statewide standardized score distribution on the state assessment. We do this in each state, separately for each grade, year, and subject. Appendix A, available in the online version of the journal, reviews the HETOP procedure. Nonetheless, the method of constructing district-level score means and standard deviations is not central to the linking methods we discuss. The following methods are applicable whenever district means and standard deviations are available with corresponding standard errors.

Fitting the HETOP model to ED*Facts* data yields estimates of each district's mean test score, where the means are expressed relative to the state's student-level population mean of 0 and standard deviation of 1, within each grade, year, and subject. We denote these estimated district means and standard deviations as $\hat{\mu}^{\text{state}}_{dygb}$ and $\hat{\sigma}^{\text{state}}_{dygb}$, respectively, for district (d), year (y), grade (g), and subject (b). The HETOP model estimation procedure also provides standard errors of these estimates, denoted $se(\hat{\mu}^{\text{state}}_{dygb})$ and $se(\hat{\sigma}^{\text{state}}_{dygb})$, respectively (Reardon et al., 2016).[5]

The second step of the linking process is to estimate a linear transformation linking each state/year/grade/subject scale (standardized to a student-level mean of 0 and standard deviation of 1—the scale of $\hat{\mu}^{\text{state}}_{dygb}$) to its corresponding NAEP distribution. Recall that we have estimates of NAEP means and standard deviations at the state (denoted s) level, denoted by $\hat{\mu}^{\text{naep}}_{sygb}$ and $\hat{\sigma}^{\text{naep}}_{sygb}$, respectively, as

well as their standard errors. To obtain estimates of these parameters in grades (3, 5, 6, and 7) and years (2010 and 2012) in which NAEP was not administered, we interpolate and extrapolate linearly.[6] First, within each NAEP-tested year, 2009, 2011, and 2013, we interpolate between Grades 4 and 8 to Grades 5, 6, and 7 and extrapolate to Grade 3. Next, for all Grades 3 through 8, we interpolate between the NAEP-tested years to estimate parameters in 2010 and 2012. We illustrate this below for means, and we apply the same approach to standard deviations.[7] Note that this is equivalent to interpolating between years first and then interpolating and extrapolating to grades:

$$\hat{\mu}^{\text{naep}}_{sygb} = \hat{\mu}^{\text{naep}}_{sy4b} + \frac{g-4}{4}(\hat{\mu}^{\text{naep}}_{sy8b} - \hat{\mu}^{\text{naep}}_{sy4b}), \quad \text{for g} \in \{3,\ 5,\ 6,\ 7\};\ \text{y} \in \{2009,\ 2011,\ 2013\};\ \text{and}\ \forall\ s, b,$$

$$\hat{\mu}^{\text{naep}}_{sygb} = \frac{1}{2}(\hat{\mu}^{\text{naep}}_{s[y-1]gb} + \hat{\mu}^{\text{naep}}_{s[y+1]gb}), \quad \text{for g} \in \{3,\ 4,\ 5,\ 6,\ 7,\ 8\};\ \text{y} \in \{2010, 2012\};\ \text{and}\ \forall\ s, b. \tag{1}$$

We evaluate the validity of linking to interpolated NAEP grades and years explicitly later in this article.

Because the estimated district test score moments $\hat{\mu}^{\text{state}}_{dygb}$ and $\hat{\sigma}^{\text{state}}_{dygb}$ are expressed on a state scale with mean 0 and unit variance, the estimated mapping of the standardized test scale in state (s), year (y), grade (g), and subject (b) to the NAEP scale is given by Equation 2. Given $\hat{\mu}^{\text{state}}_{dygb}$, this mapping yields an estimate of the of the district average performance on the NAEP scale, denoted by $\widehat{\hat{\mu}^{\text{naep}}_{dygb}}$. Given this mapping, the estimated standard deviation, on the NAEP scale, of scores in district (d), year (y), grade (g), and subject (b) is given by Equation 3:

$$\widehat{\hat{\mu}^{\text{naep}}_{dygb}} = \hat{\mu}^{\text{naep}}_{sygb} + \hat{\mu}^{\text{state}}_{dygb} \cdot \hat{\sigma}^{\text{naep}}_{sygb}, \tag{2}$$

$$\widehat{\hat{\sigma}^{\text{naep}}_{dygb}} = \hat{\sigma}^{\text{state}}_{dygb} \cdot \hat{\sigma}^{\text{naep}}_{sygb}. \tag{3}$$

The intuition behind Equation 2 is straightforward: Districts that belong to states with relatively high NAEP averages, $\hat{\mu}^{\text{naep}}_{sygb}$, should be placed higher on the NAEP scale. Within states, districts that are high or low relative to their state (positive and negative on the standardized state scale) should be relatively high or low on the NAEP scale in proportion to that state's NAEP standard deviation, $\hat{\sigma}^{\text{naep}}_{sygb}$.

From Equations 2 and 3, we can derive the (squared) standard errors of the linked means and standard deviations for noninterpolated grades and years, incorporating the imprecision from the estimates of state and NAEP means and standard deviations. In these derivations, we assume that the linking assumption is met. Later in this article, we relax this assumption and provide expressions for the standard errors of the linked means that include the linking error. For

simplicity in these derivations, we assume $\hat{\mu}^{\text{naep}}_{sygb}$ and $\hat{\sigma}^{\text{naep}}_{sygb}$ are independent random variables,[8] which yields:

$$\text{var}(\widehat{\hat{\mu}^{\text{naep}}_{dygb}}) = \text{var}(\hat{\mu}^{\text{naep}}_{sygb}) + (\hat{\sigma}^{\text{naep}}_{sygb})^2\text{var}(\hat{\mu}^{\text{state}}_{dygb}) + (\hat{\mu}^{\text{state}}_{dygb})^2\text{var}(\hat{\sigma}^{\text{naep}}_{sygb}) + \text{var}(\hat{\sigma}^{\text{naep}}_{sygb})\text{var}(\hat{\mu}^{\text{state}}_{dygb}); \quad (4)$$

$$\text{var}(\widehat{\hat{\sigma}^{\text{naep}}_{dygb}}) = \text{var}(\hat{\sigma}^{\text{state}}_{dygb})\left[\text{var}(\hat{\sigma}^{\text{naep}}_{sygb}) + (\hat{\sigma}^{\text{naep}}_{sygb})^2\right] + (\hat{\sigma}^{\text{state}}_{dygb})^2\text{var}(\hat{\sigma}^{\text{naep}}_{sygb}). \quad (5)$$

The linking Equations 2 and 3 and the standard error formulas (Equations 4 and 5) here are accurate under the assumption that there is no linking error—the assumption that the average performance of students in any given district relative to the average performance of students in their state would be the same on the NAEP and state assessments. This is a strong and untested assumption. We next provide a set of validation analyses intended to assess the accuracy of this assumption. We then provide modifications of the standard error formulas here that take the linking error into account.

Validation Checks and Results

The linking method we use here, on its own, is based on the untested assumption that districts' distributions of scores on the state accountability tests have the same relationship to one another (i.e., the same relative means and standard deviations) as they would if the NAEP assessment were administered in lieu of the state test. Implicit in this assumption is that differences in the content, format, and testing conditions of the state and NAEP tests do not differ in ways that substantially affect aggregate relative distributions. This is, on its face, a strong assumption.

Rather than assert that this assumption is valid, we empirically assess it using the patchwork reporting and administration of district results by NAEP and NWEA. We do this in several ways. First, for the districts participating in the NAEP TUDA assessments over these years, we compare $\widehat{\hat{\mu}^{\text{naep}}_{dygb}}$—the estimated district mean based on our linking method—to $\hat{\mu}^{\text{naep}}_{dygb}$—the mean of NAEP TUDA scores from the district. This provides a direct validation of the linking method, since the TUDA scores are in the metric that the linking method attempts to recover but are not themselves used in any way in the linking process. We repeat this linkage for demographic subgroups to assess the population invariance of the link.

Second, we assess the correlation of our linked district estimates with district mean scores on the NWEA MAP tests. This provides the correlation across a larger sample of districts. However, the NWEA MAP test has a different score scale, so it does not provide direct comparability with the NAEP scale that is the target of our linking.

Third, for the relevant TUDA districts, we assess whether within-district differences in linked scores across grades and cohorts correspond to those differences observed in the NAEP data. That is, we assess whether the linking provides accurate measures of changes in scores across grades and cohorts of students, in addition to providing accurate means in a given year.

Fourth, we conduct a set of validation exercises designed to assess the validity of the interpolation of the NAEP scores in non-NAEP years and grades. For all of these analyses, we present evidence regarding the district means; corresponding results for the standard deviations are in the appendices.

Validation Check 1: Recovery of TUDA Means

The NAEP TUDA data provide estimated means and standard deviations on the actual "naep" scale, $\hat{\mu}^{\text{naep}}_{dygb}$ and $\hat{\sigma}^{\text{naep}}_{dygb}$ for large urban districts in 2009 and 20 in 2011 and 2013. For these particular large districts, we can compare the NAEP means and standard deviations to their linked means and standard deviations. For each district, we obtain discrepancies $\widehat{\hat{\mu}^{\text{naep}}_{dygb}} - \hat{\mu}^{\text{naep}}_{dygb}$ and $\widehat{\hat{\sigma}^{\text{naep}}_{dygb}} - \hat{\sigma}^{\text{naep}}_{dygb}$. If there were no sampling or measurement error in these estimates, we would report the average of these discrepancies as the bias and would report the square root of the average squared discrepancies as the root mean square error (RMSE). We could also report the observed correlation between the two as a measure of how well the linked estimates align linearly with their reported TUDA values. However, because of the imprecision in both the NAEP TUDA and linked estimates the RMSE will be inflated and the correlation will be attenuated as measures of recovery. Instead, we report measurement error–corrected RMSEs and correlations that account for imprecision in both the linked and TUDA parameter estimates.

To estimate the measurement error–corrected bias, RMSE, and correlation in a given year, grade, and subject, we fit the model below using the sample of districts for which we have both estimates $\widehat{\hat{\mu}^{\text{naep}}_{dygb}}$ and $\hat{\mu}^{\text{naep}}_{dygb}$ (or $\widehat{\hat{\sigma}^{\text{naep}}_{dygb}}$ and $\hat{\sigma}^{\text{naep}}_{dygb}$ as the case may be; the model is the same for the means or standard deviations):

$$
\begin{aligned}
&\hat{\mu}_{idygb} = \alpha_{0dygb}(\text{LINKED}_i) + \alpha_{1dygb}(\text{TUDA}_i) + e_{idygb},\\
&\alpha_{0dygb} = \beta_{00} + u_{0dygb},\\
&\alpha_{1dygb} = \beta_{10} + u_{1dygb},\\
&e_{idygb} \sim N(0, \omega^2_{idygb});\ \boldsymbol{u}_{dygb} \sim \text{MVN}(0, \boldsymbol{\tau}),
\end{aligned}
\tag{6}
$$

where i indexes source (linked or NAEP TUDA test), ω^2_{idygb} is the estimated sampling variance (the squared standard error) of $\hat{\mu}_{idygb}$ (which we treat as known), and $\boldsymbol{\tau} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}$ is the variance–covariance matrix of the linked and TUDA parameter values which must be estimated. Given the model

TABLE 1.
Recovery of NAEP TUDA Means Following State-Level Linkage of State Test Score Distributions to the NAEP Scale, Measurement Error Adjusted

				Recovery		
Subject	Grade	Year	*n*	RMSE	Bias	Correlation
Reading	4	2009	17	3.95	2.12	.96
		2011	20	3.69	1.25	.96
		2013	20	2.62	0.20	.98
	8	2009	17	2.92	1.12	.95
		2011	20	2.20	0.63	.97
		2013	20	3.62	1.67	.93
Math	4	2009	17	6.09	4.10	.93
		2011	20	4.97	2.60	.94
		2013	20	3.60	1.46	.95
	8	2009	14	5.21	3.40	.95
		2011	17	3.77	2.09	.96
		2013	14	4.54	1.47	.94
Average		2009	14-17	4.70	2.69	.95
		2011	17-20	3.79	1.64	.96
		2013	14-20	3.66	1.20	.95
		Reading	17-20	3.23	1.17	.96
		Math	14-20	4.77	2.52	.95
		All	14-20	4.07	1.84	.95
Subgroup average		Male	14-20	4.14	1.84	.97
		Female	14-20	3.95	1.70	.98
		White	11-19	3.89	0.66	.98
		Black	13-19	4.11	1.80	.96
		Hispanic	12-20	4.07	2.08	.94

Source. Authors' calculations from ED*Facts* and NAEP TUDA Expanded Population Estimates data.
Note. RMSE and bias are measured in NAEP scale score points. Estimates are based on Equation 7 in text. Subgroup averages are computed from a model that pools across grades and years within subject (like Equation 9 in text); the subject averages are then pooled within subgroup. NAEP = National Assessment of Educational Progress; TUDA = Trial Urban District Assessment.

estimates, we estimate the bias, $\hat{B} = \hat{\beta}_{00} - \hat{\beta}_{10}$, and $\widehat{\mathrm{RMSE}} = [\hat{B}^2 + \mathbf{b}\hat{\boldsymbol{\tau}}\mathbf{b}']^{1/2}$ where $\mathbf{b} = [\,1 \quad -1\,]$ is a 1×2 design matrix. Finally, we estimate the correlation of α_{0dygb} and α_{1dygb} as $\hat{r} = \frac{\hat{\tau}_{01}}{\sqrt{\hat{\tau}_{00}\hat{\tau}_{11}}}$.[9]

Table 1 reports the results of these analyses in each subject, grade, and year in which we have TUDA estimates (see online Table A1 for the corresponding table for standard deviations). Although we do not show the uncorrected estimates here, we note that the measurement error corrections have a negligible impact on bias and reduce the (inflated) RMSE by around 8% on average. On average, the

linked estimates overestimate actual NAEP TUDA means by roughly 1.8 points on the NAEP scale or around .05 of a standard deviation unit, assuming an NAEP scale standard deviation of 35 (NAEP standard deviations vary from roughly 30–40 across subjects, years, and grades). The bias is slightly greater in earlier years and in mathematics.

This positive bias indicates that the average scores of students in the TUDA districts are systematically higher in the statewide distribution of scores on the state accountability tests than on the NAEP test. This leads to a higher-than-expected NAEP mapping. Table 1 also shows that the average estimated precision-adjusted correlation (disattenuated to account for the imprecision in the observed means) is .95 (note that the simple unadjusted correlation is .94; measurement error in the means is relatively minor relative to the true variation in the means of the TUDA districts). Figure 1 shows scatterplots of the estimated linked means versus the observed TUDA means, separately for grades and subjects, with the identity lines displayed as a reference.

Note that under a linear linking such as Equation 2, our definition of bias implies that the weighted average bias, among all districts within each state, and across all states, is 0 by design. If we had all districts, the bias in Table 1 would be 0; it is not 0 because Table 1 summarizes the bias for only the subset of NAEP urban districts for which we have scores. The RMSE similarly describes the magnitude of error (the square root of average squared error) for these districts and may be larger or smaller than the RMSE for other districts in the state.

We review here four possible explanations for discrepancies between a district's average scores on the state accountability test and on the NAEP assessments. These are not meant to be exhaustive explanations; they illustrate possible substantive reasons for linking error and variance. First, the population of students assessed in the two instances may differ. For example, a positive discrepancy may result if the target district excluded low-scoring students from state tests but not from NAEP. If this differential exclusion were greater in the target district, on average, than in other districts in the state, the target district would appear higher in the state test score distribution than it would in the NAEP score distribution, leading to a positive discrepancy between the district's linked mean score and its NAEP mean scores. Likewise, a positive discrepancy would result if the NAEP assessments excluded high-scoring students more in the TUDA assessment than in the statewide assessment or if there were differential exclusion of high-scoring students in other districts on the state test relative to the target district and no differential exclusion on NAEP. In other words, the discrepancies might result from a target district's scores being biased upward on the state test or downward on the NAEP assessment relative to other districts in the state and/or from other districts' scores being biased downward on the state test or upward on the NAEP assessment relative to the target district.

Second, the discrepancies may result from differential content in NAEP and state tests. If a district's position in the state distribution of skills/knowledge

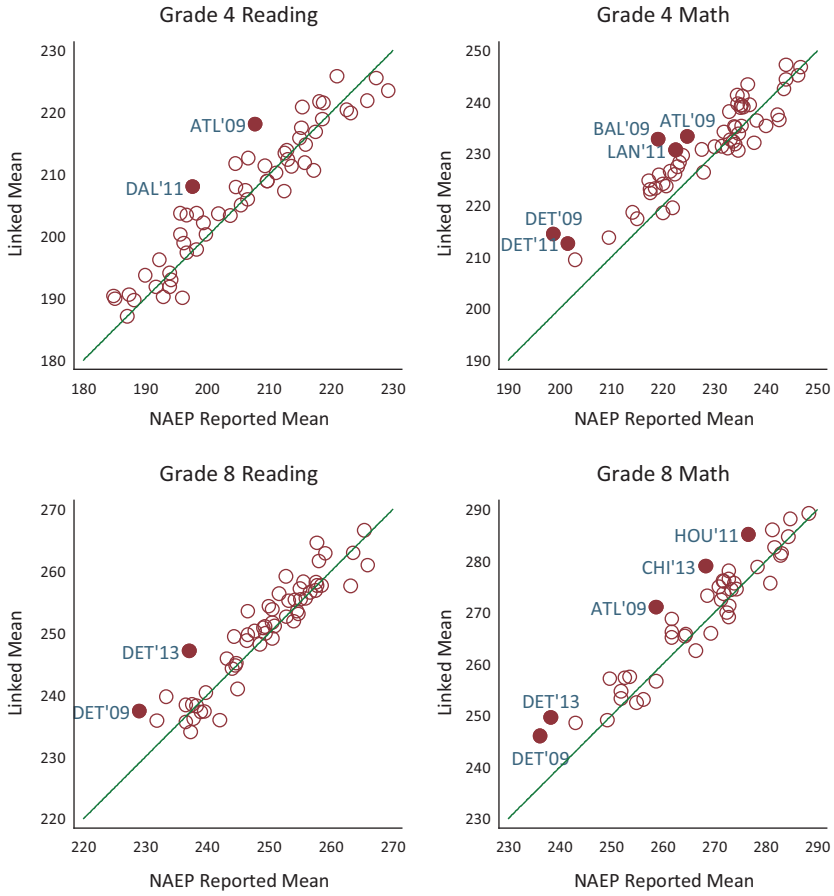

FIGURE 1. *Comparison of reported means from National Assessment of Educational Progress (NAEP) Trial Urban District Assessment and NAEP-linked state test score distributions, Grades 4 and 8, reading and mathematics, in 2009, 2011, and 2013. DAL = Dallas; ATL = Atlanta; DET = Detroit; BAL = Baltimore; LAN = Los Angeles; CHI = Chicago; HOU = Houston. District–years with a greater than 8-point discrepancy are labeled.*

measured by the state test does not match its position in the statewide distribution of skills measured by the NAEP assessment, the linked scores will not match those on NAEP. The systematic positive discrepancies in Table 1 and Figure 1 may indicate that students in the TUDA districts have disproportionately higher true skills in the content areas measured by their state tests than the NAEP assessments relative to other districts in the states. In other words, if large districts are better than other districts in their states at teaching their students the

specific content measured by state tests, relative to their effectiveness in teaching the skills measured by NAEP, we would see a pattern of positive discrepancies like that in Table 1 and Figure 1.

Third, relatedly, students in the districts with a positive discrepancy may have relatively high motivation for state tests over NAEP compared to other districts. Fourth, the bias evident in Table 1 and Figure 1 may indicate differential score inflation or outright cheating. For example, some of the largest positive discrepancies among the 20 TUDA districts illustrated in Figure 1 are in Atlanta in 2009, where there was systematic cheating on the state test in 2009 (Wilson, Bowers, & Hyde, 2011). The discrepancies in the Atlanta estimates are substantially smaller (commensurate with other large districts) in 2011 and 2013, after the cheating had been discovered. In this way, we see that many possible sources of bias in the linking are sources of bias with district scores on the state test itself rather than problems with the linking per se.

We also address the population invariance of the linking (e.g., Dorans & Holland, 2000; Kolen & Brennan, 2014) by reporting the average direction and magnitude (RMSE) of discrepancies, $\hat{\mu}^{\widehat{\text{naep}}}_{dygb} - \hat{\mu}^{\text{naep}}_{dygb}$, for selected gender and racial/ethnic subgroups in Table 1. The number of districts is lower in some grade–year cells due to insufficient subgroup samples in some districts.[10] The RMSEs are only slightly larger for subgroups than the RMSE for all students, and bias is similar in magnitude for all groups. We conclude from these comparable values that the linking functions recover NAEP district means similarly, on average, across subgroups.

Validation Check 2: Association With NWEA MAP Means

The NWEA MAP test is administered in thousands of school districts across the country. Because the MAP tests are scored on the same scale nationwide, district average MAP scores can serve as a second audit test against which we can compare the linked scores. As noted previously, in most tested districts, the number of student test scores is very close to the district's enrollment in the same subject, grade, and year. For these districts, we estimate means and standard deviations on the scale of the MAP test. The scale differs from that of NAEP, so absolute discrepancies are not interpretable. However, strong correlations between linked district means and standard deviations and those on MAP represent convergent evidence that the linking is appropriate.

We calculate disattenuated correlations between the observed MAP means deviations and both the HETOP estimated means (prior to linking them to the NAEP scale) and the linked means from Equation 2. The improvement from the correlation of MAP means and HETOP estimates to the correlation of MAP means and NAEP-linked estimates is due solely to the move from the "state" to the "$\widehat{\text{naep}}$" scale, shifting all districts within each state according to NAEP performance.

TABLE 2.
Precision-Adjusted Correlations of Linked District Means With NWEA MAP District Means Before and After State-Level Linkage of State Test Score Distributions to the NAEP Scale

				Precision-Adjusted Correlations	
Subject	Grade	Year	*N*	With HETOP Estimates	With Linked Estimates
Reading	4	2009	1,139	.90	.95
		2011	1,472	.87	.93
		2013	1,843	.92	.95
	8	2009	959	.84	.91
		2011	1,273	.87	.91
		2013	1,597	.88	.92
Math	4	2009	1,128	.86	.93
		2011	1,467	.82	.90
		2013	1,841	.87	.93
	8	2009	970	.83	.93
		2011	1,279	.84	.92
		2013	1,545	.87	.95
Average		2009	4,196	.86	.93
		2011	5,491	.85	.91
		2013	6,826	.88	.94
All Years			16,513	.87	.93

Source. Authors' calculations from ED*Facts* and NWEA MAP data.
Note. Sample includes districts with reported NWEA MAP scores for at least 90% of students. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress; NAEP = National Assessment of Educational Progress; HETOP = heteroskedastic ordered probit.

Table 2 shows that correlations between the linked district means and MAP district means are .93 on average when adjusting for imprecision (see online Table A2 for the corresponding table for standard deviations). This is larger than the average correlation of .87 between the MAP means and the (unlinked) HETOP estimates. Figure 2 shows a bubble plot of district MAP scores on linked scores for Grade 4 mathematics in 2009 as an illustration of the data underlying these correlations. Note that the points plotted in Figure 2 are means estimated with imprecision. The observed (attenuated) correlations are generally .03 to .10 points lower than their disattenuated counterparts.

Validation Check 3: Association of Between-Grade and Between-Cohort Trends

An additional assessment of the extent to which the linked state district means match the corresponding NAEP district means compares not just the means in a given grade and year but the within-district differences in means across grades

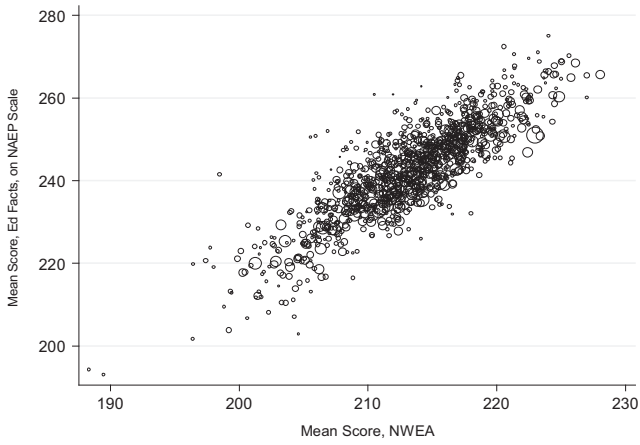

FIGURE 2. *Example of an association between linked means and Northwest Evaluation Association Measures of Academic Progress means, Grade 4 math, 2009. Correlation of .87; precision-adjusted correlation of .93. Bubble size corresponds to district enrollment.*

and years. If the discrepancies evident in Figure 1 are consistent across years and grades within a district, then the linked state estimates will provide accurate measures of the within-district trends across years and grades, even when there is a small bias in in the average means.

To assess the accuracy of the across-grade and across-year differences in linked mean scores, we use data from the 20 TUDA districts from the grades and years in which we have both linked means and corresponding means from NAEP. We do not use the NAEP data from interpolated years and grades in this model. We fit the same model for both means and standard deviations and separately by subject. For each model, we fit precision-weighted random coefficients models of this form:

$$
\begin{aligned}
\hat{\mu}_{idygb} &= \alpha_{0dygb}(\text{LINKED}_i) + \alpha_{1dygb}(\text{TUDA}_i) + e_{idygb},\\
\alpha_{0dygb} &= \beta_{00d} + \beta_{01d}(\text{year}_{dygb} - 2011) + \beta_{02d}(\text{grade}_{dygb} - 6) + u_{0dygb},\\
\alpha_{1dygb} &= \beta_{10d} + \beta_{11d}(\text{year}_{dygb} - 2011) + \beta_{12d}(\text{grade}_{dygb} - 6) + u_{1dygb},\\
\beta_{00d} &= \gamma_{00} + v_{00d},\\
\beta_{01d} &= \gamma_{01} + v_{01d},\\
\beta_{02d} &= \gamma_{02} + v_{02d},\\
\beta_{10d} &= \gamma_{10} + v_{10d},\\
\beta_{11d} &= \gamma_{11} + v_{11d},\\
\beta_{12d} &= \gamma_{12} + v_{12d},\\
e_{idygb} &\sim N(0, \boldsymbol{\omega}^2_{idygb});\ \boldsymbol{u}_{dygb} \sim \text{MVN}(0, \boldsymbol{\sigma}^2);\ ;\ \boldsymbol{v}_d \sim \text{MVN}(0, \boldsymbol{\tau}),
\end{aligned}
\tag{7}
$$

where i indexes source (linked or NAEP TUDA test) and ω^2_{idygb} is the sampling variance of $\hat{\mu}_{idydb}$ (which we treat as known and set equal to the square of the

estimated standard error of $\hat{\mu}_{idydb}$). The vector $\mathbf{\Gamma} = \{\gamma_{00}, \ldots, \gamma_{12}\}$ contains the average intercepts, year slopes, and grade slopes (in the second subscript, 0, 1, and 2, respectively) for the linked values and the target values (in the first subscript, 0 and 1, respectively). The differences between the corresponding elements of $\mathbf{\Gamma}$ indicate average bias (i.e., the difference between γ_{00} and γ_{10} indicates the average deviation of the linked means and the NAEP TUDA means, net of district-specific grade and year trends). Unlike Table 1, where we estimated bias separately for each year and grade and descriptively averaged them, the bias here is estimated by pooling over all years and grades of TUDA data, with district random effects. If the linking were perfect, we would expect this to be 0.

The matrix of random parameters $\boldsymbol{\tau}$ includes, on the diagonal, the between-district variances of the average district means and their grade and year trends; the off-diagonal elements are their covariances. From $\boldsymbol{\tau}$, we can compute the correlation between the within-district differences in mean scores between grades and years. The correlation $\text{corr}(v_{01d}, v_{11d})$, for example, describes the correlation between the year-to-year trend in district NAEP scores and the trend in the linked scores. Likewise the correlation $\text{corr}(v_{02d}, v_{12d})$ describes the correlation between the Grade 4 through 8 differences in district NAEP scores and the corresponding difference in the linked scores. Finally, the correlation $\text{corr}(v_{00d}, v_{10d})$ describes the correlation between the NAEP and linked intercepts in the model—that is, the correlation between linked and TUDA mean scores. This correlation differs from that shown in Table 1 because the former estimates the correlation separately for each grade and year; the model in Equation 7 estimates the correlation from a model in which all years and grades are pooled.

Table 3 shows the results of fitting this model separately by subject to the district means (see online Table A3 for the corresponding table for standard deviations). When comparing the linked estimates to the NAEP TUDA estimates, several patterns are evident. First, the estimated correlation of the TUDA and linked intercepts is .98 (for both math and reading), and the bias in the means (the difference in the estimated intercepts in Table 3) is small and not statistically significant. The linked reading means are, on average, 1.1 points higher (*SE* of the difference is 3.0; not statistically significant) than the TUDA means, and the linked mathematics means are, on average, 2.4 points higher (*SE* of the difference is 3.3, not statistically significant) than the TUDA means. These are, not surprisingly, similar to the average bias estimated from each year and grade separately and shown in Table 1.

Second, the estimated average linked and TUDA grade slopes ($\hat{\gamma}_{02}$ and $\hat{\gamma}_{12}$, respectively) are nearly identical to one another in both math and reading. The estimated bias in grade slopes ($-.04$ in reading and $-.10$ in math) is only 1% as large as the average grade slope. The implied RMSE from the model is .56 in

TABLE 3.
Estimated Comparison of Linked and TUDA District Means, Pooled Across Grades and Years, by Subject

Parameter	Reading	Math
Linked ED*Facts* parameters		
Intercept (γ_{00})	228.53***	250.59***
	(2.00)	(2.10)
Year (γ_{01})	0.91***	0.43*
	(0.17)	(0.19)
Grade (γ_{02})	10.81***	9.58***
	(0.27)	(0.29)
TUDA parameters		
Intercept (γ_{10})	227.41***	248.14***
	(2.12)	(2.49)
Year (γ_{11})	1.03***	0.91***
	(0.10)	(0.11)
Grade (γ_{12})	10.84***	9.68***
	(0.22)	(0.17)
L2 intercept *SD*s: linked (σ_0)	2.51	2.66
L2 intercept *SD*s: TUDA (σ_1)	0.82	1.26
Correlation: L2 residuals	1.00	0.36
L3 intercept *SD*s: linked (τ_1)	8.87	9.27
L3 intercept *SD*s: TUDA (τ_4)	9.79	11.10
L3 year slope *SD*s: linked (τ_2)	—	—
L3 year slope *SD*s: TUDA (τ_5)	—	—
L3 grade slope *SD*s: linked (τ_3)	1.06	1.03
L3 grade slope *SD*s: TUDA (τ_6)	0.93	0.61
Correlation: L3 intercepts	0.98	0.98
Correlation: L3 year slopes	—	—
Correlation: L3 grade slopes	0.85	0.98
Reliability L3 intercept: linked	0.98	0.98
Reliability L3 year slope: linked	—	—
Reliability L3 grade slope: linked	0.76	0.73
Reliability L3 intercept: TUDA	1.00	1.00
Reliability L3 year slope: TUDA	—	—
Reliability L3 grade slope: TUDA	0.87	0.72
N: observations	228	204
N: districts	20	20

Note. Estimates are based on Equation 9 in text. The Level 3 random errors on the year slope were not statistically significant and so were dropped from the model. L2 = "Level 2"; L3 = "Level 3." NAEP = National Assessment of Educational Progress; TUDA = Trial Urban District Assessment.
Source. Authors' calculations from ED*Facts* and NAEP TUDA Expanded Population Estimates data.
***$p \leq .001$; **$p \leq .01$; *$p \leq .05$.

reading and .46 in math, roughly 5% of the average grade slope.[11] The estimated correlation of the TUDA and linked grade slopes is .85 for reading and .98 for math. Finally, the reliability of the grade slopes of the linked estimates is .76 in reading and .73 in math.[12] Together these indicate that the linked estimates provide unbiased estimates of the within-district differences across grades and that these estimates are precise enough to carry meaningful information about between-grade differences.

Third, there is little or no variation in the year trends in the TUDA districts; for both math and reading, the estimated variation of year trends is small and not statistically significant. As a result, neither the TUDA nor the linked estimates provide estimates of trends across years that are sufficiently reliable to be useful (in models not shown, we estimate the reliabilities of the TUDA year trends to be .28 and .53 and of the linked year trends to be .45 and .72 in reading and math, respectively). As a result, we dropped the random effects on the year trends and do not report in Table 3 estimates of the variance, correlation, or reliability of the year trends.

Validation Check 4: Recovery of Estimates Under Interpolation Between Years and Grades

Using the interpolated state means and standard deviations in Equation 1 for the linking establishes an assumption that the linkage recovers district scores that would have been reported in years 2010 and 2012 and Grades 3, 5, 6, and 7. Although we cannot assess recovery of linkages in interpolated grades with only Grades 4 and 8, we can check recovery for an interpolated year, specifically, 2011, between 2009 and 2013. By pretending that we do not have 2011 NAEP state data, we can assess performance of our interpolation approach by comparing linked estimates to actual 2011 TUDA results. For each of the TUDAs that participated in both 2009 and 2013, we interpolate, for example,

$$\begin{aligned}\hat{\mu}^{\text{naep}\prime}_{s2011gb} &= \frac{1}{2}(\hat{\mu}^{\text{naep}}_{s2009gb} + \hat{\mu}^{\text{naep}}_{s2013gb}),\\ \hat{\sigma}^{\text{naep}\prime}_{s2011gb} &= \frac{1}{2}(\hat{\sigma}^{\text{naep}}_{s2009gb} + \hat{\sigma}^{\text{naep}}_{s2013gb}).\end{aligned} \tag{8}$$

Applying Equations 2 through 5, we obtain estimates, for example, $\widehat{\hat{\mu}^{\text{naep}}_{d2011gb}}'$, and we compare these to actual TUDA estimates from 2011. We estimate bias and RMSE for discrepancies $\widehat{\hat{\mu}^{\text{naep}}_{d2011gb}}' - \hat{\mu}^{\text{naep}}_{d2011gb}$ using the model from Validation Check 1. Table 4 shows results in the same format as Table 1 (see online Table A4 for the corresponding table for standard deviations). We note that the average RMSE of 3.8 and bias of 1.4 in Table 4 are approximately the same as the average RMSE of 3.8 and bias of 1.6 shown for 2011 in Table 1. Note that the

TABLE 4.
Recovery of Reported 2011 NAEP TUDA Means Following State-Level Linkage of State Test Score Distributions to an NAEP Scale Interpolated Between 2009 and 2013, Measurement Error Adjusted

				Recovery		
Subject	Grade	Year	*N*	RMSE	Bias	Correlation
Reading	4	2011	20	3.78	0.89	.95
	8	2011	20	2.14	1.47	.99
Math	4	2011	20	4.67	2.25	.94
	8	2011	14	3.81	1.66	.96
Average		Reading	20	3.07	1.18	.97
		Math	14-20	4.26	1.95	.95
		All	14-20	3.72	1.57	.96

Note. Using NAEP Expanded Population Estimates. Adjusted correlations account for imprecision in linked and target estimates.

interpolations we use in our proposed linking are those between observed scores that are only 2 years apart rather than 4 years apart as in the validation exercise here. The 2-year interpolations should be more accurate than the 4-year interpolation, which itself is accurate enough to show no degradation in our recovery of estimated means. We conclude that the between-year interpolation of state NAEP scores adds no appreciable error to the linked estimates for TUDA districts.

We next investigate the viability of interpolation by comparing correlations of linked district estimates with MAP scores at different degrees of interpolation. Some grade–year combinations need no interpolation, others are singly interpolated, and others are doubly interpolated. Table 5 shows that, on average, precision-adjusted correlations between linked NAEP means and MAP means are almost identical across different degrees of interpolation, around .93 (see online Table A5 for the corresponding table for standard deviations). This lends additional evidence that interpolation adds negligible aggregate error to recovery between years as well as grades.

Quantifying the Uncertainty in Linked Estimates

The validation analyses above suggest that the linked estimates correspond quite closely to their target values on average. But, as is evident in Table 1 and Figure 1, the degree of discrepancy varies among districts. NAEP and state assessments do not locate districts identically within states, implying linking error in the estimates. In this section, we provide a framework for quantifying

TABLE 5.
Precision-Adjusted Correlations Between NWEA MAP District Means and NAEP-Linked Estimates

Subject	Grade	2009	2010	2011	2012	2013
Reading	3	*.95*	**.94**	*.93*	**.93**	*.94*
	4	.95	*.94*	.93	*.94*	.95
	5	*.94*	**.94**	*.93*	**.93**	*.94*
	6	*.92*	**.94**	*.92*	**.93**	*.93*
	7	*.92*	**.93**	*.92*	**.92**	*.92*
	8	.91	*.91*	.91	*.91*	.92
Math	3	.91	**.89**	*.91*	**.91**	*.91*
	4	.93	*.92*	.90	*.92*	.93
	5	*.91*	**.90**	*.92*	**.91**	*.93*
	6	*.93*	**.93**	*.94*	**.93**	*.94*
	7	*.94*	**.95**	*.95*	**.95**	*.95*
	8	.93	*.93*	.92	*.94*	.95
No interpolation (plain text)		.93				
Single interpolation (italics)		*.93*				
Double interpolation (boldface)		**.93**				

Note. Linked using NAEP Expanded Population Estimates. Sample includes districts with reported NWEA MAP scores for at least 90% of students. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress; NAEP = National Assessment of Educational Progress. All correlations are statistically significant, $p < .001$.

the magnitude of this linking error. This in turn allows us to construct approximate standard errors for cross-state comparisons based on the linking.

We begin by rewriting Equation 2, omitting the *ygb* subscripts and using superscripts *n* and *s* in place of "naep" and "state" for parsimony. Here, we distinguish the linked mean for a district *d* (estimated using Equation 2 and denoted here as $\hat{\mu}_d^{\text{link}}$) from the mean NAEP score we would observe if all students in a district took the NAEP assessment (this is the target parameter, denoted here as μ_d^n):

$$\begin{aligned}
\hat{\mu}_d^{\text{link}} &= \hat{\mu}_s^n + \hat{\mu}_d^s \cdot \hat{\sigma}_s^n, \\
&= (\mu_s^n + \nu_s^n) + (\mu_d^s + \nu_d^s) \cdot (\sigma_s^n + w_s^n), \\
&= (\mu_s^n + \mu_d^s \sigma_s^n) + (\mu_d^s w_s^n + \sigma_s^n \nu_d^s + \nu_d^s w_s^n + \nu_s^n), \\
&= \mu_d^{\text{link}} + (\mu_d^s w_s^n + \sigma_s^n \nu_d^s + \nu_d^s w_s^n + \nu_s^n), \\
&= \mu_d^n + \delta_d + (\mu_d^s w_s^n + \sigma_s^n \nu_d^s + \nu_d^s w_s^n + \nu_s^n).
\end{aligned} \tag{9}$$

Equation 10 shows that the estimated linked mean has two sources of error: linking error (δ_d) and sampling/estimation error (the term in parentheses) that results from error in the estimated state mean and standard deviation on the NAEP assessment and from estimation error in the district mean on the state

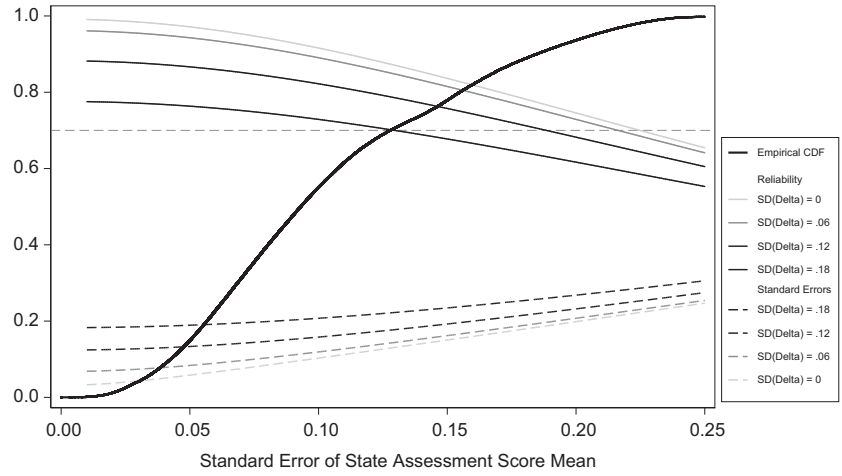

FIGURE 3. *Reliabilities and standard errors (in National Assessment of Educational Progress [NAEP] national standard deviation units) of linked means on the NAEP scale across a range of specifications.*

assessment. The variance of the sampling/estimation error term is given by Equation 4. The variance of the linking error is not known for the full population, but we can illustrate its magnitude using the RMSEs for TUDAs reported in Table 1. The average root mean squared linking error in Table 1 is about 4 points on the NAEP scale. Because the standard deviation of national NAEP scores within a grade, year, and subject is typically 28 to 38 points, the RMSE is roughly 4/33≈.12 student-level standard deviation units. To the extent that the RMSE of the linking errors across the TUDA districts is representative of the standard deviation of the linking errors in the full population, we can then approximate the standard deviation of δ_d as .12.

We can now compute approximate standard errors of the linked estimates that take into account linking error as well as sampling and estimation error. The variance of the linked mean will be

$$\mathrm{var}(\hat{\mu}_d^{\mathrm{link}}) = \mathrm{var}(\delta_d) + (\mu_d^s)^2\mathrm{var}(w_s^n) + (\sigma_s^n)^2\mathrm{var}(v_d^s) + \mathrm{var}(v_d^s)\cdot\mathrm{var}(w_s^n) + \mathrm{var}(v_s^n). \quad (10)$$

We have estimates of σ_s^n, $\mathrm{var}(v_s^n)$, and $\mathrm{var}(w_s^n)$ from NAEP; we have estimates of μ_d^s and $\mathrm{var}(v_d^s)$ from the state assessment data, and we have an approximation of $\mathrm{var}(\delta_d)$ from Table 1. Using appropriate estimates of these parameters, we can compute standard errors of the linked assessments from Equation 10.

How large is the linking error relative to the other sources of error in the linked estimate? We can get a sense of this from Figure 3, which shows the standard error of the linked estimate as a function of the magnitude of the standard error of the state assessment mean, $SD(v_d^s)$, and the linking error variance, $SD(\delta_d)$. The

standard errors of the state assessment means in our data range from roughly .01 to .25, with a median of roughly .10 (the cumulative distribution function of the distribution of standard errors is shown in Figure 3 for reference). The figure includes four lines describing the standard error of the linked means under four different assumptions about the magnitude of the linking error: $SD(\delta_d) \in \{0, .06, .12, .18\}$. The scenario $SD(\delta_d) = 0$ corresponds to the assumption that there is no linking error (an unrealistic assumption but useful as a reference). The TUDA analyses in Table 1 suggest a value of $SD(\delta_d) = .12$; we include higher and lower values for $SD(\delta_d)$ in order to describe the plausible range. The other terms in Equation 10 are held constant at values near their empirical medians: $\mu_d^s = 0$, $\sigma_s^n = .98$, $\text{var}(v_s^n) = .001$, and $\text{var}(w_s^n) = .0004$. The computed standard errors are very insensitive to these terms across the full range of their empirical values in the NAEP and state assessment data we use.

Figure 3 shows that the standard errors of the linked estimates are meaningfully larger when we take into account the linking error than when we assume it is 0, but this is more true for districts with small standard errors on the state assessment. In terms of standard errors, linking error plays a larger role when other sources of error are small.

Figure 3 also describes the implied reliability of the linked estimates. We estimate the standard deviation of the district means (excluding measurement error) across all districts to be roughly .34 national student-level standard deviations. Using this, we compute the reliability as $r = \frac{.34}{.34 + \text{var}(\hat{\mu}_d^{\text{link}})}$. Under the assumption of linking error with $SD(\delta_d) = .12$, the reliability of the linked means is lower than if we assume no linking error (as expected). However, it is still about .70 for the roughly 91% of districts with state standard errors less than .19. This suggests that, even in the presence of linking error, there is enough signal in the linked estimates to be broadly useful for distinguishing among districts.

We can also compute standard errors of the difference between two districts' means. The formula will differ for districts in the same state versus different states. For two districts in the same state

$$\text{var}\left(\hat{\mu}_{d1}^{\text{link}} - \hat{\mu}_{d2}^{\text{link}}\right) \approx 2\text{var}(\delta) + (\mu_{d1}^s - \mu_{d2}^s)^2 \text{var}(w_s^n) + (\sigma_s^n)^2[\text{var}(v_{d1}^s) + \text{var}(v_{d2}^s)]. \quad (11)$$

For two districts in different states, however,

$$\begin{aligned}\text{var}\left(\hat{\mu}_{d1}^{\text{link}} - \hat{\mu}_{d2}^{\text{link}}\right) \approx\ & 2\text{var}(\delta) + (\mu_{d1}^s)^2 \text{var}(w_{s1}^n) + (\mu_{d2}^s)^2 \text{var}(w_{s2}^n) + (\sigma_{s1}^n)^2 \text{var}(v_{d1}^s) \\ & + (\sigma_{s2}^n)^2 \text{var}(v_{d2}^s) + \text{var}(v_{s1}^n) + \text{var}(v_{s2}^n). \quad (12)\end{aligned}$$

Assuming the sampling variances of the state means and standard deviations are the same in both states (which is approximately true in this case, given that NAEP sample sizes and standard deviations are similar among states), this is

$$\mathrm{var}\left(\hat{\mu}_{d1}^{\mathrm{link}} - \hat{\mu}_{d2}^{\mathrm{link}}\right) \approx 2\mathrm{var}(\delta) + [(\mu_{d1}^{s})^2 + (\mu_{d2}^{s})^2]\mathrm{var}(w_s^n) + (\sigma_s^n)^2[\mathrm{var}(v_{d1}^{s}) + \mathrm{var}(v_{d2}^{s})] + 2\mathrm{var}(v_s^n). \tag{13}$$

We ignore the $\mathrm{var}(v_d^s) \cdot \mathrm{var}(w_s^n)$ terms because they are very small relative to the other terms in the formula. The difference in the variance of a between-state and a within-state comparison, holding all other terms constant, will be $2\mu_{d1}^2\mu_{d2}^2\mathrm{var}(w_s^n) + 2\mathrm{var}(v_s^n)$. The same-state and different-state formulas differ because the within-state comparisons share the same sampling error in the state NAEP means and standard deviations. As a result, there is generally more uncertainty in between-state comparisons than within-state comparisons on the NAEP scale. However, the difference is generally small. Both between-state and within-state comparisons share the linking error variance $\left(2\mathrm{var}(\delta)\right)$ and the sampling/estimation error variance in the district means $\left(\mathrm{var}(v_{d1}^{s}) + \mathrm{var}(v_{d2}^{s})\right)$. In addition, $\mathrm{var}(w_s^n)$ and $\mathrm{var}(v_s^n)$ are small relative to these two sources of error in the case we examine here. The result is that the errors in within- and between-state differences in linked means are generally similar.[13]

Discussion

We present validation methods for aggregate-level linking that applies whenever some subset of groups has scores on both tests. We motivate the method with the goal of constructing a U.S.-wide district-level data set of test score means and standard deviations. We demonstrate that test score distributions on state standardized tests can be linked to a national NAEP-linked scale in a way that yields district-level distributions that correspond well—but not perfectly—to the absolute performance of TUDA districts on NAEP and the relative performance of available districts on MAP. The correlation of district-level mean scores on the NAEP-linked scale with scores on the NAEP TUDA and NWEA MAP assessments is generally high (averaging .95 and .93 across grades, years, and subjects). Nonetheless, we find some evidence that NAEP-linked estimates include some small, but systematically positive, bias in large urban districts (roughly +.05 standard deviations, on average). This implies a corresponding small downward bias for other districts in the same states, on average.

Are these discrepancies a threat to the validity of the linked estimates of district means? The answer depends on how the estimates will be used. Given evidence of the imperfect correlation and small bias, the linked estimates should not be used to compare or rank school districts' performance when the estimated means are close and when the districts are in different states. As we noted, there are several possible sources of error in a cross-state comparison, including differences in content, motivation, sampling, and inflation. Our methods cannot

identify the presence of any one type of error but do allow us to quantify the total amount of error in cross-state TUDA comparisons. This error is small relative to the distribution of test scores and the variation in average district scores. Of course, relative comparisons within states do not depend on the linking procedure, so these are immune to bias and variance that arises from the linking methods.

On the basis of these results, we believe the linked estimates are accurate enough to be used to investigate broad patterns in the relationships between average test performance and local community or schooling conditions, both within and between states. The validation exercises suggest that the linked estimates can be used to examine variation among districts and across grades within districts. It is unclear whether the estimates provide unbiased estimates of within-grade trends over time, given that there is little or no variation in the NAEP district trends over time against which to benchmark the linked trend estimates. This is true more generally even of within-grade national NAEP trends, which are often underpowered to detect true progress over shorter time spans of 2 to 4 years.

Validation methods must begin with an intended interpretation or use of scores (Kane, 2013). An operational interpretation of the linked aggregate estimates is the result of monotonic transformations of district score distributions on state tests. They are state score distributions with NAEP-based adjustments, with credit given for being in a state with relatively high NAEP performance and, for districts within the states, greater discrimination among districts when a state's NAEP standard deviation is high. Our contribution is to provide a strategy and methods for answering the essential counterfactual question: What would district results have been, had district scores on NAEP or MAP been available? When patchwork administrations of district tests are available, we can obtain direct and indirect answers to this question.

Because some combination of testing conditions, purpose, motivation, and content of NAEP and state tests differ, we find that district results do differ across tests. But our validation checks suggest that these differences are generally small relative to the variation among districts. This is evident in the high correspondence of the linked and NAEP TUDA estimates and of the linked and NWEA MAP estimates. This suggests that our set of estimated NAEP-linked district test score results may be useful in empirical research describing and analyzing national variation in local academic performance. When data structures with patchwork administrations of tests are available in other U.S. and international testing contexts, our strategy and methods are a road map to not only link scores at the aggregate level but to validate interpretations and uses of linked scores.

Authors' Note

Some of the data used in this article were provided by the National Center for Education Statistics (NCES). This article would not have been possible without the assistance of

Ross Santy, Michael Hawes, and Marilyn Seastrom, who facilitated access to the ED*Facts* data. This article benefited substantially from ongoing collaboration with Erin Fahle, Ken Shores, and Ben Shear. The opinions expressed here are our own and do not represent views of NCES, Northwest Evaluation Association, the Institute of Education Sciences, the Spencer Foundation, the Russell Sage Foundation, the Bill and Melinda Gates Foundation, the Overdeck Family Foundation, the William T. Grant Foundation, or the U.S. Department of Education.

Acknowledgment

We are grateful to Yeow Meng Thum at Northwest Evaluation Association (NWEA), who provided the NWEA data used in some analyses.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research described here was supported by grants from the Institute of Education Sciences (R305D110018), the Spencer Foundation, the Russell Sage Foundation, the Bill and Melinda Gates Foundation, the Overdeck Family Foundation, and the William T. Grant Foundation.

Notes

1. Our data and methods are more comprehensive than those used in the *Global Report Card* (GRC; Greene & McGee, 2011; Greene & Mills, 2014; http://globalreportcard.org/). First, we provide grade-specific estimates (by year), allowing for estimates of measures of progress. Second, instead of the statistical model we describe below (Reardon, Shear, Castellano, & Ho, 2016), which leverages information from three cut scores in each grade, the GRC uses only one cut score and aggregates across grades. This assumes that stringency is the same across grades and that district variances are equal. Third, our methods allow us to provide standard errors for our estimates. Fourth, we provide both direct and indirect validation checks for our linkages.
2. From 2009 to 2013, 63% of 3,060 state(51)–grade(6)–subject(2)–year(5) reliability coefficients were available. Consistent with Reardon and Ho (2015), the reported reliabilities have a mean of .905 and a standard deviation of .025. Missing reliabilities were imputed as predicted values from a linear regression of reliability on state–grade–subject and state–year fixed effects. The residuals from the model have a standard deviation of .010. This suggests that imputation errors based on the model have a standard deviation of .010, which is very small compared to the mean value of the reliabilities

(.905). As a result, imputation errors in reliability are not likely to be consequential.

3. In 2009, the 17 districts are Atlanta, Austin, Baltimore, Boston, Charlotte, Chicago, Cleveland, Detroit, Fresno, Houston, Jefferson County, Los Angeles, Miami, Milwaukee, New York City, Philadelphia, and San Diego. Albuquerque, Dallas, and Hillsborough County joined in 2011 and 2013. Washington, D.C., is not included for validation, as it has no associated state for linking. California districts (and Texas districts in 2013) did not have a common Grade 8 state mathematics assessment, so the California and Texas districts lack a linked district mean for Grade 8 mathematics.
4. Note that the correlation of Expanded Population Estimates and regular National Assessment of Educational Progress (NAEP) estimates are near unity; as a result, our central substantive conclusions are unchanged if we use the regular NAEP estimates in the linking.
5. Note that, because there is measurement error in the state accountability test scores, estimates of μ_{dygb} and σ_{dygb} that are standardized based on the observed score distribution will be biased estimates of the means and standard deviations expressed in terms of the standardized true score distribution (the means will be biased toward 0; the standard deviations will be biased toward 1). Before linking, we adjust $\hat{\mu}^{\text{state}}_{dygb}$ and $\hat{\sigma}^{\text{state}}_{dygb}$ to account for measurement error using the classical definition of reliability as the ratio of true score variance over observed score variance. We adjust the means and their standard errors by dividing them by the square root of the state test score reliability (ρ) in the relevant year, grade, and subject. We adjust the standard deviations and their standard errors by multiplying them by $\sqrt{\frac{\hat{\sigma}^2+\rho-1}{\rho\hat{\sigma}^2}}$. After these adjustments, $\hat{\mu}^{\text{state}}_{dygb}$ and $\hat{\sigma}^{\text{state}}_{dygb}$ are expressed in terms of the standardized distribution of true scores within the state. We do not adjust NAEP state means and standard deviations, as NAEP estimation procedures account for measurement error due to item assignment to examinees (Mislevy, Muraki, & Johnson, 1992).
6. Interpolation relies on some comparability of NAEP scores across grades. Vertical linking was built into NAEP's early design via cross-grade blocks of items (administered to both fourth and eighth graders) in 1990 in mathematics and in 1992 in reading (Thissen, 2012). These cross-grade blocks act as the foundation for the bridge spanning Grades 4 and 8. At around that time, the National Assessment Governing Board that sets policy for NAEP adopted the position that, as Haertel (1991) describes, "NAEP should employ within-age scaling whenever feasible" (p. 2). Thissen notes that there have been few checks on the validity of the cross-grade scales since that time. One exception is a presentation by McClellan, Donoghue, Gladkova, and Xu (2005) who tested whether subsequent vertical linking would have made a difference on the reading assessment. They concluded that "the

current cross-grade scale design used in NAEP seems stable to the alternate design studied" (p. 37). We do not use interpolation between grades to explore growth trajectories on an absolute scale but rather identify relative position of districts, which are very consistent across grades. Our fourth validation check provides evidence that this interpolation is justified.

7. Note that the sampling variances of the interpolated means and standard deviations will be functions of the sampling variances of the noninterpolated values. For example,

$$\mathrm{var}(\hat{\mu}^{\mathrm{naep}}_{sygb}) = \left(\frac{8-g}{4}\right)^2 \mathrm{var}(\hat{\mu}^{\mathrm{naep}}_{sy4b}) + \left(\frac{g-4}{4}\right)^2 \mathrm{var}(\hat{\mu}^{\mathrm{naep}}_{sy8b}),$$

$$\text{for g} \in \{3,\ 5,\ 6,\ 7\};\ \text{y} \in \{2009,\ 2011,\ 2013\};\ \text{and}\ \forall\ s, b,$$

$$\mathrm{var}(\hat{\mu}^{\mathrm{naep}}_{sygb}) = \frac{1}{4}[\mathrm{var}(\hat{\mu}^{\mathrm{naep}}_{s[y-1]gb}) + \mathrm{var}(\hat{\mu}^{\mathrm{naep}}_{s[y+1]gb})],$$

$$\text{for g} \in \{3,\ 4,\ 5,\ 6,\ 7,\ 8\};\ \text{y} \in \{2010, 2012\};\ \text{and}\ \forall\ s, b.$$

8. This is not strictly true, since $\hat{\mu}^{\mathrm{naep}}_{sygb}$ and $\hat{\sigma}^{\mathrm{naep}}_{sygb}$ are estimated from the same sample. However, the NAEP samples are large within each state–year–grade–subject, so the covariance of the estimated means and standard deviations is very small relative to other sources of sampling variance in Equation 4.
9. Note that this model assumes the errors e_{idygb} are independent within each district–grade–year–subject. The error in the NAEP Trial Urban District Assessment (TUDA) estimate arises because of sampling variance (because the NAEP assessment was given to only a random sample of students in each TUDA district). The error in the linked estimate arises because of (a) error in the estimated district mean score and (b) sampling error in the NAEP estimates of the state mean and standard deviation (see Equation 4). The error in the estimated state mean arises from the fact that the heteroskedastic ordered probit model estimates the mean score from coarsened data not from sampling variance (because the state assessments include the full population of students); the error in the NAEP state mean and standard deviation arises from sampling variance in the state NAEP samples. Both the coarsening error and the state sampling error are independent of the sampling error in the NAEP district mean estimate.
10. Our model for subgroups pools across grades (4 and 8) and years (2009, 2011, and 2013) to compensate for smaller numbers of districts in some grade–year cells. We describe this model in Validation Check 3. On average, across grades and years, the results are similar to a model that does not use pooling. We also calculate bias and root mean square error (RMSE) for Asian student populations but do not report them due to small numbers of districts: 5 to 10 per cell. However, bias and RMSE of linked district

estimates were higher for Asians, suggesting caution against conducting a separate linkage for Asian students.

11. We compute the RMSE of the grade slope from the model estimates as follows. Let $\hat{C} = \hat{\gamma}_{02} - \hat{\gamma}_{12}$ be the bias in the grade slopes; then, the RMSE of the grade slope will be $\widehat{\text{RMSE}} = [\hat{C}^2 + \mathbf{d}\hat{\boldsymbol{\tau}}^2\mathbf{d}']^{1/2}$, where $\mathbf{d} = [0\ 0\ 1\ 0\ 0 - 1]$.
12. The reliability of the Level 3 slopes and intercepts is computed as described in Raudenbush and Bryk (2002).
13. To illustrate this, suppose we assume a low value for the linking error variance (say $\text{var}(\delta) = .0036$, one quarter of what is implied by the RMSE of the NAEP TUDA means). Let us compare two districts with small standard errors (say $\text{var}(\hat{\mu}_{d1}) = \text{var}(\hat{\mu}_{d1}) = .0025$, which is smaller than 90% of districts; see Figure 3). If both districts have estimated means two standard deviations from their state means (so $\hat{\mu}_{d1} = \hat{\mu}_{d2} = 2$, an extreme case), then Equation 11 indicates that the error in a within-state comparison will have variance of .0122, while Equation 13 indicates that a between-state comparison will have a modestly larger error variance of .0174. In most comparisons (where the error variances of the estimated district means are larger, where the district means are not so far from their state averages, or when the linking error variance is larger), the difference in the error variance of between- and within-state comparisons will be much smaller.

References

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). *Mapping state proficiency standards onto NAEP scales: 2005-2007 (NCES 2010-456)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Braun, H., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales. Statistics for social and behavioral sciences* (pp. 313–338). New York, NY: Springer.

Braun, H., Zhang, J., & Vezzu, S. (2010). An investigation of bias in reports of the National Assessment of Educational Progress. *Educational Evaluation and Policy Analysis*, *32*, 24–43.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Gewertz, C. (2015). PARCC restructures, allows states to customize test. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/curriculum/2015/11/parcc_allows_states_to_customize_test.html

Greene, J. P., & McGee, J. M. (2011). *George W. Bush Institute global report card: Technical Appendix*. Report prepared for the George W. Bush Institute. Retrieved

from http://www.globalreportcard.org/docs/AboutTheIndex/Global-Report-Card-Technical-Appendix-9-28-11.pdf

Greene, J. P., & Mills, J. N. (2014). *George W. Bush Institute global report card 3.0: Technical Appendix*. Report prepared for the George W. Bush Institute. Retrieved from http://www.bushcenter.org/stateofourcities/data/GWBI_GRC_TechnicalAppendix.pdf

Haertel, E. H. (1991). *Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational progress*. Washington, DC: National Center for Educational Statistics. Retrieved from https://files.eric.ed.gov/fulltext/ED404367.pdf

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, *17*, 267–321.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking*. New York, NY: Springer-Verlag.

McClellan, C. A., Donoghue, J. R., Gladkova, L., & Xu, X. (2005). *Cross-grade scales in NAEP: Research and real-life experience*. Longitudinal Modeling of Student Achievement. Maryland Assessment Research Center for Education Success. College Park, MD: University of Maryland. Retrieved from http://marces.org/conference/Longitudinal/Mcclellan.ppt

McLaughlin, D. (2005). *Properties of NAEP full population estimates* (Unpublished report). American Institutes for Research. Retrieved from http://www.schoolata.org/Portals/0/uploads/reports/NSA_T1.5_FPE_Report_090205.pdf

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154.

National Institute of Statistical Sciences. (2009). *NISS/NESSI task force on full population estimates for NAEP* (Technical report #172). Retrieved from http://www.niss.org/sites/default/files/technical_reports/tr172.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, *40*(2): 158–189.

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2016). *Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data*. Retrieved from https://cepa.stanford.edu/sites/default/files/wp16-02-v201601.pdf

Reardon, S. F., Yun, J. T., & Eitle, T. M. (1999). *The changing context of school segregation: Measurement and evidence of multi-racial metropolitan area school segregation, 1989-1995*. Paper presented at the annual meeting of the American Educational Research Association. Montreal, Canada.

Stroub, K. J., & Richards, M. P. (2013). From resegregation to reintegration: Trends in the racial/ethnic segregation of metropolitan public schools, 1993–2009. *American Educational Research Journal*, *50*, 497–531.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287–312). New York, NY: Springer.

Thissen, D. (2012). *Validity issues involved in cross-grade statements about NAEP results. NAEP validity studies panel*. Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2009). *Race to the top executive summary*. Washington, DC. Retrieved from https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2015). ED*Facts Submission System User Guide V11.2* (SY 2014-2015). Washington, DC: ED Facts. Retrieved from http://www.ed.gov/EDFacts

U.S. Department of Education. (n.d.). *NAEP data explorer*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences.

Whitehurst, G. J., Chingos, M. M., & Gallaher, M. R. (2013). *Do school districts matter?* Washington, DC: Brookings Institution.

Wilson, R. E., Bowers, M. J., & Hyde, R. L. (2011). *Special investigation into test tampering in Atlanta's school system*. Retrieved from https://ia801000.us.archive.org/32/items/215252-special-investigation-into-test-tampering-in/215252-special-investigation-into-test-tampering-in.pdf

Authors

SEAN F. REARDON is a professor of poverty and inequality in education and professor (by courtesy) of sociology at Stanford University, 520 Galvez Mall Drive, Stanford, CA 94305; email: sean.reardon@stanford.edu. His research focuses on the causes, patterns, trends, and consequences of social and educational inequality; the effects of educational policy on educational and social inequality; and applied statistical methods for education research.

DEMETRA KALOGRIDES is a research associate at the Center for Education Policy Analysis at Stanford University, 520 Galvez Mall Drive, Stanford, CA 94305; email: dkalo@stanford.edu. Her research interests are educational inequality, teacher and principal labor markets, and quantitative research methods.

ANDREW D. HO is a professor of education at Harvard Graduate School of Education, 455 Gutman Library, 6 Appian Way, Cambridge, MA 02138; email: andrew_ho@gse.harvard.edu. His research addresses educational accountability metrics that support inferences about proficiency, growth, college readiness, course completion, and value added.

Manuscript received June 21, 2017
First revision received July 30, 2018
Second revision received April 2, 2019
Accepted August 2, 2019